特別研究報告

題目

深層学習に基づく行動に着目したシーン抽出手法

指導教員
村田 正幸 教授

報告者
山西 宏平

2015 年 2 月 13 日

大阪大学 基礎工学部 情報科学科

平成 26 年度 特別研究報告

深層学習に基づく行動に着目したシーン抽出手法

山西 宏平

内容梗概

複雑なモデルを表現できる多階層のニューラルネットワークを用いた深層学習と呼ばれる学習手法が、注目を集めている。深層学習では、プレトレーニングと呼ばれる教師なし学習で、各ニューロンの接続の重みを決めた上で、教師あり学習を行うことにより、複雑なモデルの学習を行う。深層学習は、様々な分野で応用されるようになっており、特に画像や動画像の認識への適用は広く研究されている。それらの研究では、画像や動画像ファイル全体に対して識別・分類の処理を行い、正確に画像や動画に写っている事象を識別・分類できることが示されている。しかしながら、これらの従来研究では、動画ファイル全体や、識別したい時点の前後のフレームの情報を用いて動画の識別を行うことを対象としている。それに対して、監視カメラにおける異常検出や長い動画像データからの必要な場面の映像の抽出など、動画像内の各時点のシーンを把握することが重要な動画像識別のアプリケーションも存在する。深層学習を用いて、動画像の各時刻のシーンを識別する方法としては、動画像中の現在から N フレーム前までの連続したフレームを入力とし、識別したいシーンの分類結果を出力とするようなニューラルネットワークを構成することが考えられる。しかしながら、各時点のシーンの識別には、動画像内の動きを認識する必要があり、動きの認識に必要な時間分の全フレームを入力としたニューラルネットワークを構成すると、入力ユニット数が多く複雑なニューラルネットワークが構成される。その結果、識別したいシーンに対する学習データ数が少ない場合は、十分な学習ができずに、正確な識別ができなくなる可能性がある。

そこで、本報告では、現在から N フレーム前までのフレームのうち、サンプリングした少数のフレームを入力としたニューラルネットワークを用いて、動画像のシーンを識別する手法を提案する。提案手法では、現在のシーンの識別に重要であると考えられる、直近のフレームは短い間隔でサンプリングを行い、過去にいくほどサンプリング間隔を広くする。サンプリングされた入力は、畳み込みニューラルネットワークの入力として用いる。これにより、シーンを識別するのに必要な少数の入力と出力を対応付けるニューラルネットワークを構成することができ、識別対象のシーンに対する学習データが少ない場合であっても、正確な識別が期待できる。

本報告では、映像内の人物の動作をシーンのラベルとしてつけ、シーンの切り替わりのある動画像を用いて、提案手法の評価を行った。評価では、サンプリング間隔を過去に行くほど 1 ずつ増やしながら、サンプリングした 8 フレームを入力としたニューラルネットワークを構成した提案手法、連続した直近 8 フレームを入力として用いた手法、連続した 29 フレームを入力とした手法を比較した。いずれの入力に対しても、8 階層の畳み込みニューラルネットワークを構成し、提案手法と連続した 29 フレームを入力とした手法では 33040 個、連続した 8 フレームを入力とした手法では 38440 個のサンプルを用いて学習を行った。その結果、連続したフレームを入力として用いた手法では、いずれの時刻においても、60%以上の識別率を達成することはできなかったのに対して、提案手法では、シーン切り替わり後、1 秒以降であれば、95%以上の精度でシーンの識別を行うことができることが分かった。

主な用語

深層学習、ニューラルネットワーク、畳み込み、特徴量抽出、フレーム

目 次

表 目 次

図 目 次

1　はじめに

近年、多階層のニューラルネットワークを用いた機械学習手法である、深層学習が様々な分野で注目されている。階層型ニューラルネットワークは、階層数を増やすことにより、表現できるモデルの自由度が増え、複雑なモデルを表現できるようになる。しかしながら、多階層のニューラルネットワークを学習させることは困難であった。この問題に対して、多階層の各層において、教師なし学習を繰り返すことによるプレトレーニングを実行した後に、教師ありデータを用いてニューラルネットワーク全体の調整を行う学習法 [1] が提案され、多階層のニューラルネットワークを効率的に学習することができるようになり、様々な分野で深層学習を応用した研究が進められるようになってきた。

画像認識は、深層学習の応用として盛んに研究が進められている分野の一つである。画像認識とは、画像の特徴量抽出を行い、抽出をした結果を機械学習と組み合わせて認識を行うことである。これまで、画像の特徴量の抽出手法として様々な手法が提案されてきた [2-4]。それに対して、深層学習を用いた手法 [5, 6] では、画像の生データをニューラルネットワークの入力として用い、特徴量の抽出の手順から、機械学習により学習を行う。深層学習を用いた学習手法は、正確な画像認識を行うことができることが示されている [7]。

動画像の認識に深層学習を用いる手法も検討が進められている。文献 [8] では、動画ファイルの分類に対して、深層学習の適用が検討されており、各フレームの画像のみを入力に用いた学習よりも、複数フレームのデータを入力に用いた畳み込みニューラルネットワークを用いた学習が、より正確に動画像ファイルに映っている事象の識別ができることを明らかにしている。

しかしながら、これらの従来研究では、あらかじめ定められた範囲に含まれるフレーム全体を用いて動画の識別を行うことを対象としている。それに対して、監視カメラにおける異常検出や長い動画像データからの必要な場面の映像の抽出など、動画像内の各時点において発生した事象を把握することが重要な画像識別のアプリケーションも存在する。深層学習を用いて、動画像の各時刻のシーンを識別する方法としては、動画像中の現在から N フレーム前までの連続したフレームを入力とし、識別したいシーンの分類結果を出力とするようなニューラルネットワークを構成することが考えられる。しかしながら、各時点のシーンを識別には、動画像内の動きを認識する必要があり、動きの認識に必要な時間分の全フレームを入力としたニューラルネットワークを構成すると、入力ユニット数が多く複雑なニューラルネットワークが構成される。その結果、識別したいシーンに対する学習データ数が少ない場合は、十分な学習ができずに、正確な識別ができなくなる可能性がある。

そこで、本報告では、現在から N フレーム前までのフレームのうち、サンプリングした少数のフレームを入力としたニューラルネットワークを用いて、動画像のシーンを識別する

手法を提案する。提案手法では、現在のシーンの識別に重要であると考えられる、直近のフレームは短い間隔でサンプリングを行い、過去にいくほどサンプリング間隔を広くする。サンプリングされた入力は、畳み込みニューラルネットワークの入力として用いる。これにより、シーンを識別するのに必要な少数の入力と出力を対応付けるニューラルネットワークを構成することができ、識別対象のシーンに対する学習データが少ない場合であっても、正確な識別が期待できる。

以降、2 章で深層学習の概要及び関連研究について紹介する。その後、本報告で用いた深層学習に基づくシーン検出手法について 3 章で述べ、4 章でそれらの比較評価を行い、提案手法の有用性について考察を行う。最後に 5 章で、まとめについて述べる。

2　深層学習

深層学習とは、多階層のニューラルネットワークを用いた機械学習手法である。本節では、深層学習で用いられるニューラルネットワークの概要について紹介をしたのち、深層学習におけるニューラルネットワークの学習法について説明する。

2.1　ニューラルネットワーク

2.1.1　ニューラルネットワークの構造

階層型ニューラルネットワークは図 1 のような構造を持つ。図中の○は一つのニューロンを表す。各ニューロンは一つの入力 x に対し一つの出力 y を返す性質があり、活性化関数 $y = f(x)$ によって x に対する y が決定される。接続された 2 つの層について、入力する側の層は下位層、入力を受ける側の層は上位層と呼ばれる。活性化関数には sigmoid 関数 $f(x) = \frac{1}{1+e^{-x}}$ や rectified linear 関数 $f(x) = \max(0, x)$ などが使われている。上位層のニューロン j への入力 x_j は式 (1) に示すように下位層の接続されたニューロンからの出力 $y_i (i = 1, ..., m)$ にニューロン i, j 間に設定された重み $w_{i,j}$ を掛けたものの総和にバイアス b_j と呼ばれる値を加算したものとなる。

$$x_j = b_j + \sum_{i=1}^{m} y_i w_{ij} \tag{1}$$

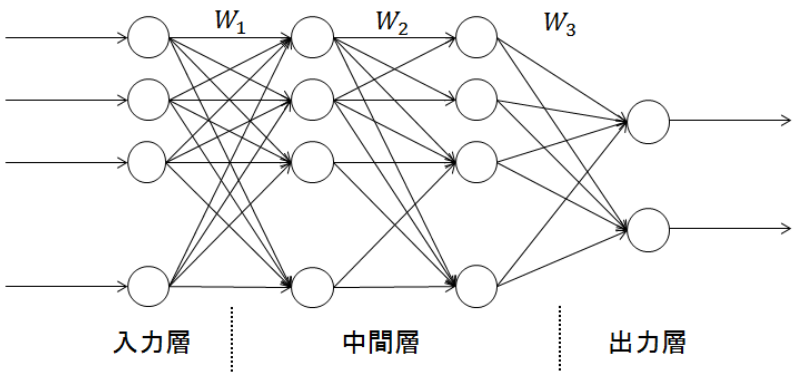

図 1: 階層型ニューラルネットワーク

階層型ニューラルネットワークでは、各ニューロンがこの計算をし、上位のニューロンに伝達をするということを繰り返すことで、全体として見ると入力層に入力された値の組に対して出力層まで計算を行い、出力層で出力を行う関数として動作する。この関数の動作は、ニューラルネットワークの重み $w_{i,j}$ とバイアス B_j で定義される。そのため、ニューラルネットワークを学習させる際には、入力に対応する適切な出力をするように、$w_{i,j}$、b_j を調整する。

ニューラルネットワークは、画像識別などのクラス分類問題によく適用される。クラス分類問題では、入力に対して、適切なクラスを出力として回答することが求められる。クラス分類問題にニューラルネットワークを適用する際には、分類先のクラス数と同数のニューロンを出力層に配置する。そして、その出力層のニューロンの値にソフトマックス関数を適用する。

ソフトマックス関数は n 個のニューロンをもつある層の $i(=1,...,n)$ 番目のニューロンが持つ値を a_i としたとき、そのニューロンの値を入力としたときに式 (2) で表される $f(a_i)$ を出力する関数である。

$$f(a_i) = \frac{\exp(a_i)}{\sum_{j=1}^{n} \exp(a_j)} \tag{2}$$

出力層全体のニューロンにソフトマックス関数を適用すると、適用後の層の全てのニューロンの持つ値が、総和が 1 になるように正規化される。このため、$f(a_i)$ をクラス i に該当する確率として扱うことができる。

2.1.2 畳み込みニューラルネットワーク

ニューラルネットワークの応用先の一つとして、画像識別が挙げられる。画像識別では、画像データをニューラルネットワークの入力として用いる。画像データは、二次元空間に配置された画素のデータからなる。各画素は、赤の明度、緑の明度、青の明度という 3 つの値を持っている。このような空間の各位置に関する情報の組はチャネルと呼ばれる。

画像データのような空間上の複数チャネルの値からなるデータをニューラルネットの入力として用いる場合は、その空間的な位置の情報を利用した、畳み込みニューラルネットワークとよばれるニューラルネットワークが用いられる。畳み込みニューラルネットワークは、図 2 で示すように、畳み込み層とよばれる層と、プーリング層とよばれる層を中間層に配置する。畳み込み層やプーリング層の各ニューロンは、下位層のすべてのニューロンと接続するのではなく、フィルタと呼ばれる空間的な範囲内のニューロンとのみ接続する。これにより、中間層の各ニューロンは、空間的に近接した範囲内の特徴量を取り出す役割を果たす。そして、出力層において、直下の畳み込み層やプーリング層のニューロンと全結合した構成を取ることにより、畳み込み層やプーリング層で抽出された特徴量と出力の対応関係を表す。

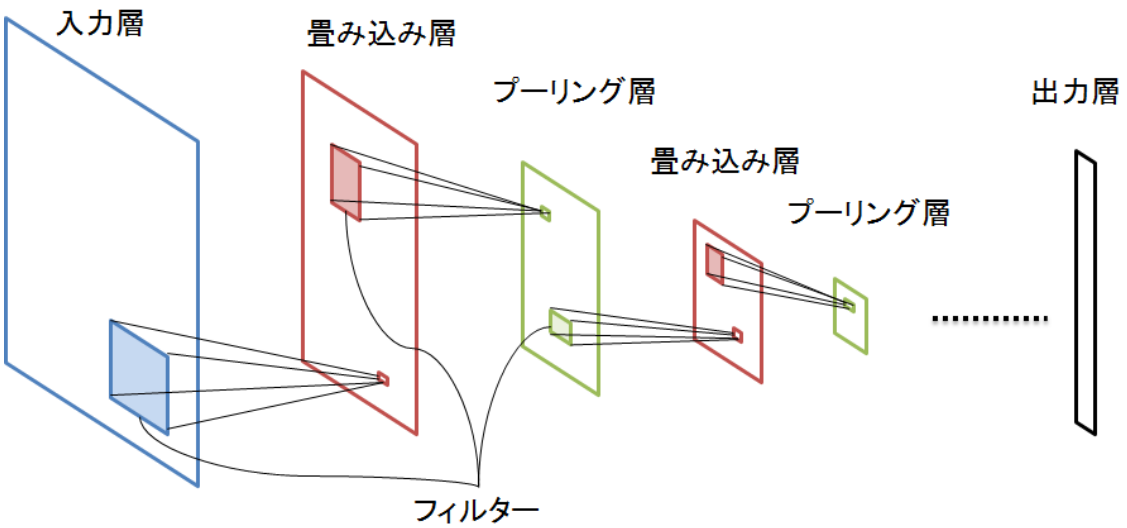

図 2: 畳み込みニューラルネットワーク

各畳み込み層は、図 3 で示すような構成をとる。図 3 では、フィルタ内の 3 × 3 の空間における 5 チャネルの値をしめす下位層の 45 個のニューロンから、1 × 1 空間における 10 チャネルの値をしめす 10 個のニューロンと接続している。この下位層 45 個のニューロンと上位層 10 個のニューロンは全結合をする。この畳み込み層を通すことにより、下位層の二次元空間の入力が、複数のより小さな範囲の二次元空間へ写像される。その結果、下位層の二次元空間の持つ特徴をより少なうニューロン数であらわすことができるようになる。

それに対して、プーリング層は、各フィルタに対して、フィルタ内のニューロンの値に対して演算を行い、1 つの値を出力する処理を行う層である。プーリングとして最も一般的な手法は領域中の最大値を取り出して代表値として集約するマックスプーリングと呼ばれる手法である。図 4 のように領域をそれぞれ切り出して各領域の強い特徴を持っている部分をそれぞれ集約することで、ニューロン数の少ないものにまとめられるだけでなく、領域内での微小な位置のずれによる影響を緩和することができるという特徴がある。

畳み込みニューラルネットワークでは、各層におけるフィルタと、畳み込み後のチャネル数の定め方が重要となる。また、フィルタについては、畳み込みを行う空間的な範囲の大きさを示すフィルタの大きさのみならず、フィルタをどれだけずつずらしながら畳み込みを行

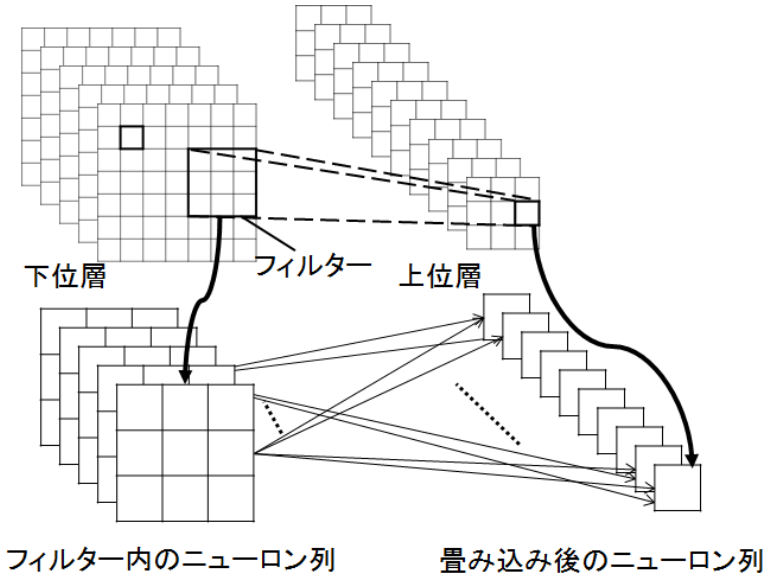

図 3: 畳み込み

うのかを示すフィルタの間隔についても定めることが必要となる。フィルタの間隔は、フィルタ一辺の長さと等しい値に設定すると、隣合うフィルタ同士で重複せずにフィルタの範囲を定めることができる。しかしながら、この場合、フィルタの境界上に重要な特徴が存在した場合、その特徴を抽出するような畳み込みが難しくなる。そのため、一般的には、フィルタの間隔は、フィルタの一辺の長さの半分程度に設定し、隣り合うフィルタ同士の範囲が重なるように設定される。また、フィルタの範囲を大きくすると、フィルタ内に含まれる特徴量も多くなる。そのため、畳み込み後のチャネル数も十分に大きな値に設定しないと、フィルタ内の特徴量が失われてしまう。そのため、フィルタの範囲にあわせて、畳み込み後のチャネル数は決める必要がある。

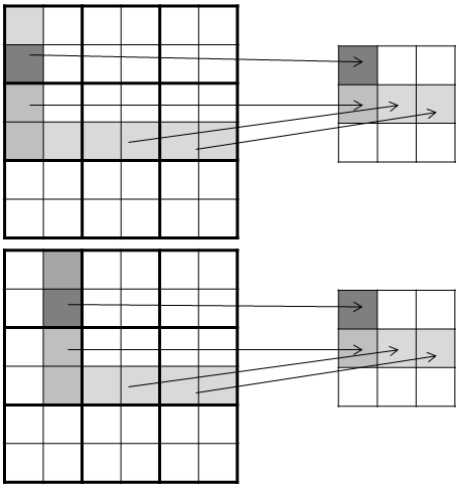

図 4: マックスプーリングによる位置ずれの無視

2.2 ニューラルネットワークによる学習方法

2.2.1 誤差逆伝搬法

ニューラルネットワークをクラス分類問題に適用した場合には、ラベル i に該当するデータを入力したとき式 (2) の値が 1 となるような出力を行うようにニューラルネットワーク中の重み $w_{i,j}$、b_i を学習する。学習の際には、正解クラスラベルと関連付けた学習サンプルを用意する。そして、ニューラルネットワークに学習サンプルを入力し、その誤差を計算、誤差が小さくなるように $w_{i,j}$、b_i の調整を行う。この学習においては、以下の式で定義される交差エントロピー誤差が用いられる。

$$C = -\sum_{i=1}^{n} q_i \log p_i$$

ただし、p_i は得られた出力、q_i は目標とする出力である。交差エントロピーは 2 つの確率分布 p, q の相違を表す尺度であり、$p = q$ のときに最小になる性質を持つ。そのため、交差エントロピーを最小化することにより、出力を目標値にすることができる。

誤差 C の最小化には勾配降下法が用いられ、その計算の効率化のために誤差逆伝搬法が用いられる。勾配降下法ではサンプルを入力したときに出力から誤差 C を得て誤差勾配 $\frac{\partial C}{\partial w_{ij}}$ を計算するとき、出力層から遠い中間層や入力層からの重みに対する誤差に対しては複数の

入れ子になった合成関数の偏微分を計算することになり、出力層から遠ざかるほど計算式が複雑になる。誤差逆伝搬法はこれを偏微分の連鎖法則を用いて計算する方法で、誤差を出力層から入力層に向けて逆向きに伝搬しながら順に重みやバイアスの更新式を計算していく。また、一度にすべての学習サンプルに対してではなく、ミニバッチと呼ばれる数百程度の学習サンプル集合に対して学習を行う、確率的勾配降下法と呼ばれるパラメータ更新法が用いられる。確率的勾配降下法では、重みとバイアスの修正は式 (3) のように行う。

$$\Delta w_{ij}^{(t)} = -\epsilon \frac{\partial C}{\partial w_{ij}^{(t)}} + \alpha \Delta w_{ij}^{(t-1)} - \epsilon \lambda w_{ij}^{(t)} \tag{3}$$

ただし $\Delta w_{ij}^{(t)}, \Delta w_{ij}^{(t-1)}$ はそれぞれ今回, 前回の重み更新時の修正量を表す。式 (3) のうち、第一項は C の偏微分によって得られた誤差勾配に学習率 ϵ を掛けたものであり、勾配下降法により誤差を削減するための $w_{i,j}$ の修正量を表す項である。確率的勾配降下法では、さらに前回の修正量の $\alpha(\sim 0.9)$ 倍で表されるモメンタムと呼ばれる項を加えることでミニバッチの選び方による偏りを抑え、現在の重みの定数倍で表される重み減衰項を加えることで重みが大きくなりすぎないようにしている。

2.2.2 深層学習による学習方法

プレトレーニング ニューラルネットワークで学習を行う際、ネットワークの各重みの初期値は乱数により決定することが一般的であったが、この初期値が適切なものでないことから多階層ニューラルネットワークでは過学習が起こっていた。過学習とは、学習サンプルに対する誤差ばかりが小さくなりそれ以外の未知のサンプルに対する誤差が小さくならないことである。深層学習において過学習を回避する方法がプレトレーニングという教師なし学習である。プレトレーニングでは、隣り合った各 2 層のネットワークを入力層側から順にそれぞれ独立して訓練を行い、特徴抽出に適した重みを得られるようにする。図 1 のようなニューラルネットワークに対しプレトレーニングを行う場合の一例として、まず入力層と、その一つ上位の中間層に対し訓練を行うために図 5 のようなニューラルネットワークを用意する。図中の 2 つの四角で囲んだ部分は同一のもので層 1、層 2 はそれぞれ図 1 の入力層とその一つ上位の中間層と同じ層であり、層 3 は層 1 と同じニューロン数を持つ、元のニューラルネットワークとは無関係の層である。このようなニューラルネットワークにおいて、入力と出力が一致するように、重み W_1 の調整を行う。次に同様に W_2 を決定する。このとき、入力は学習サンプルを層 1 に入力したときの層 2 の出力を用いる。多階層ニューラルネットワークでは、このような重みの調整を各階層で行うことにより、プレトレーニングを行う。なお、出力層に繋がるネットワークの重みは以降で述べるファインチューニングで調整すればよいためプレトレーニングを行わない。

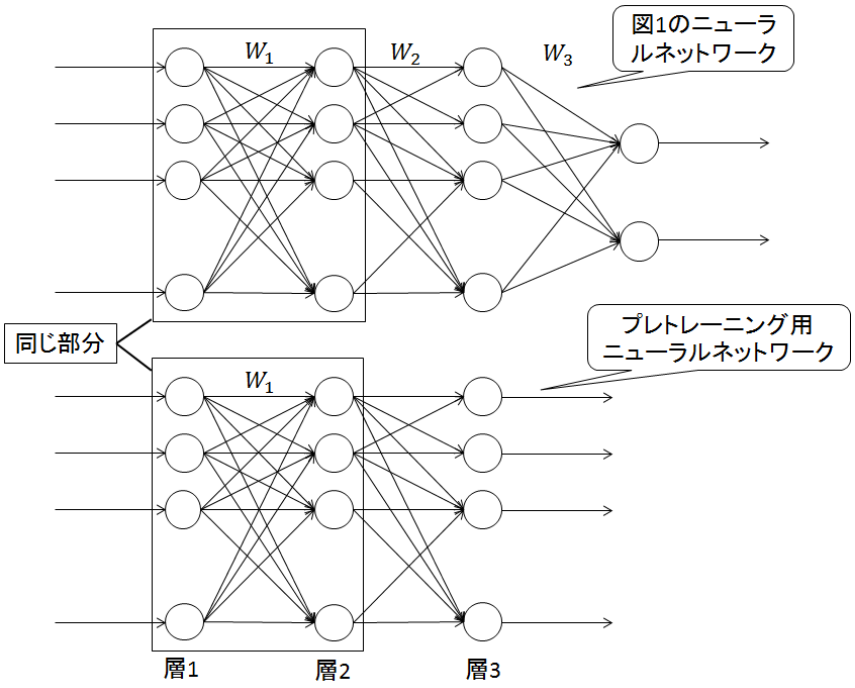

図 5: プレトレーニング用のニューラルネットワーク

ファインチューニング　プレトレーニングで得られたネットワークの重みを初期値として、学習サンプルを用いてネットワーク全体を微調整する。初期値がプレトレーニングにより最適化されていることを除き、一般的なニューラルネットワークの学習と同様に、2.2.1 項で述べたように誤差逆伝搬法により重みやバイアスの調整が行われる。

2.3　関連研究

2.3.1　深層学習を用いた画像認識

ニューラルネットワークを用いた画像認識では、入力に画像の各画素のデータを用い、中間層で特徴量の抽出を行い、分類結果を出力する。入力された画像データに対して畳み込み層を複数並べて局所的な特徴量の抽出を繰り返すニューラルネットワークの構成がよく用いられている。文献 [5] では最初の畳み込み層では 224 × 224 の画像データに対し畳み込み層、プーリング層、畳み込み層、プーリング層、3 つの畳み込み層、プーリング層、2 つの全結合層、出力層の順に層間の接続を行い、入力画像を 1000 種類に分類するニューラルネットワークが構成されている。このような畳み込みニューラルネットワークは深層学習以外の手法に比べて高い画像認識性能をもつ [7]。しかしながら、画像 1 枚分のデータを入力とする手法では一般物体の分類や物体の形の認識は高い精度で可能な反面、物の動きについて識別を行うことは難しい。

2.3.2　深層学習を用いた動画像認識

深層学習を用いた画像認識手法が確立され、高い成果を上げることが示されたため、近年、深層学習を用いて動画像を認識する場合についても検討が進められている。文献 [8] ではスポーツなどが映された動画像データセットに対しクラス分類が行われている。動画像内の一つのフレームのみを入力に用いる場合、動画像内の指定された区間の最初と最後のフレームのみを入力に用いる場合、数枚の連続したフレームを入力に用いる場合、数枚の連続したフレームの組を複数入力に用いる場合で結果が比較されており、入力データをフレーム 1 枚分のデータよりも枚数を多く、動画像中の被写体の動きの情報が増えていくほど識別の正確性が上がることが示されている。しかし、この手法では、動画像全体の情報を用いて大規模なニューラルネットワークを構成する。そのため、（１）正確に学習を行うためには、多量の学習データを必要とし、学習データが少ない場合は十分な精度での識別ができない、（２）大きな計算機資源が必要となる、という問題がある。それに対して、本報告では、動画像全体ではなく、動画像の各時刻を対象とし、その時刻で発生しているシーンを識別し、必要なシーンを抽出することを目的としている。必要なシーンを定義する学習サンプルは、多数準備することが難しい場合も考えられ、本報告では、そのような学習サンプル数が少ない場合であっても、十分な精度で識別ができるような、深層学習を適用したシーン抽出手法を提案する。

3　深層学習に基づくシーン検出手法

3.1　概要

本報告では、深層学習を応用し、動画像のうち各時刻に発生している事象を識別することを目的とする。各時刻に発生している事象を識別するには、その時刻の静止画のみでは、動画像内の人物の行動や写っている物体の移動を識別することができない。そのため、ある時刻に発生している事象を判断するには、当該時刻の画像のみならず、過去から当該時刻までのフレームの変化を入力として用いる必要がある。しかしながら、過去から当該時刻までの全フレームを入力とした学習は、入力ユニット数が多く複雑なニューラルネットワークが構成される。その結果、識別したいシーンに対する学習データ数が少ない場合は、十分な学習ができずに、正確な識別ができなくなる可能性がある。ただし、現在発生していることを把握するためには、直近の動画像データに関しては、細粒度の情報を用い細かな動作を把握することが有用であると考えられるが、過去の動画像に関しては、細粒度な情報は必要ないと考えられる。そこで、本報告では、各時刻に発生している事象を把握することを目指したニューラルネットワークの応用方法として、ニューラルネットワークへの入力に、直近は短い間隔でサンプリングしたフレーム、過去は長い間隔でサンプリングしたフレームのみを用いる手法を提案する。

3.2　ニューラルネットワークの構成

本節では、提案手法で用いるニューラルネットワークの構成を入力層、中間層、出力層にわけて説明する。

3.2.1　入力層

提案手法では、入力層の各ニューロンには、動画内のシーンを識別するのに必要なフレームの画素の値を与える。シーンの識別には、画像に写っている人物の動きを捉えることが必要となる。そのため、図 6 に示されるように、現時刻から過去にさかのぼって一定期間のフレームを入力として用いることが必要となる。ただし、範囲内の全フレームを入力対象とすると、入力ユニット数が多くなり、複雑なニューラルネットワークが構成される。その結果、識別したいシーンに対する学習データ数が少ない場合は、十分な学習ができずに、正確な識別ができなくなる可能性がある。

そのため、本報告では、過去のフレームのうち、サンプリングされた一部のフレームのみ入力として用いる。現在のシーンを識別する際には、直近のフレームに対して、過去のフ

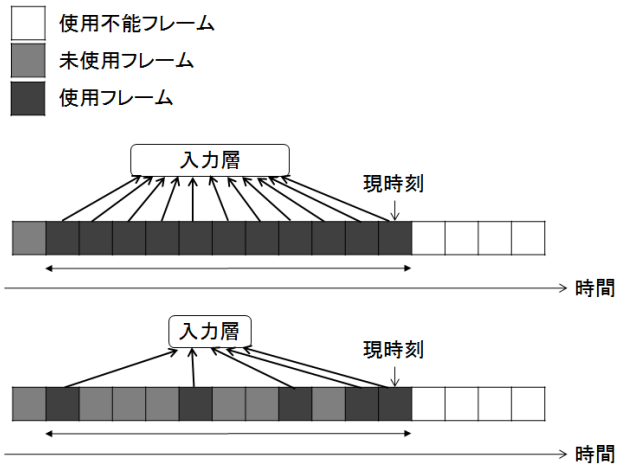

図 6: 入力に使用するフレームの選択

レームの重要度は低くなると考えられる。そのため、過去になればなるほど、サンプリングレートを下げるようにサンプリングされたフレームを用いる。提案手法では、図 6 下に示すように、最初のサンプリングを行うフレーム位置を現時刻とし、サンプリングを行うごとにそのフレーム位置を k フレーム分過去のものにずらしながら次のサンプリングを行う。k の値は初期値を 1 として、サンプリングを行うごとに 1 ずつ増加させる。このようにサンプリングを行うフレーム位置の間隔を大きくしていくことで、過去になればなるほどサンプリングレートが下がっていくようにサンプリングされたフレームを入力をすることが可能となる。

3.2.2 中間層

中間層では、畳み込み層とプーリング層を配置した畳み込みニューラルネットワークを構成する。動画像の認識においては、各画素について入力として用いられたフレーム数分の、各時刻のデータが存在する。本ニューラルネットワークでは、同一座標のデータ、全フレーム分をチャネルとして扱う。そして、図 7 のように、フィルタ内の空間に含まれる全フレームのデータを畳み込むことにより、フィルタ内での時間変化を捉えた特徴量の抽出が可能となる。

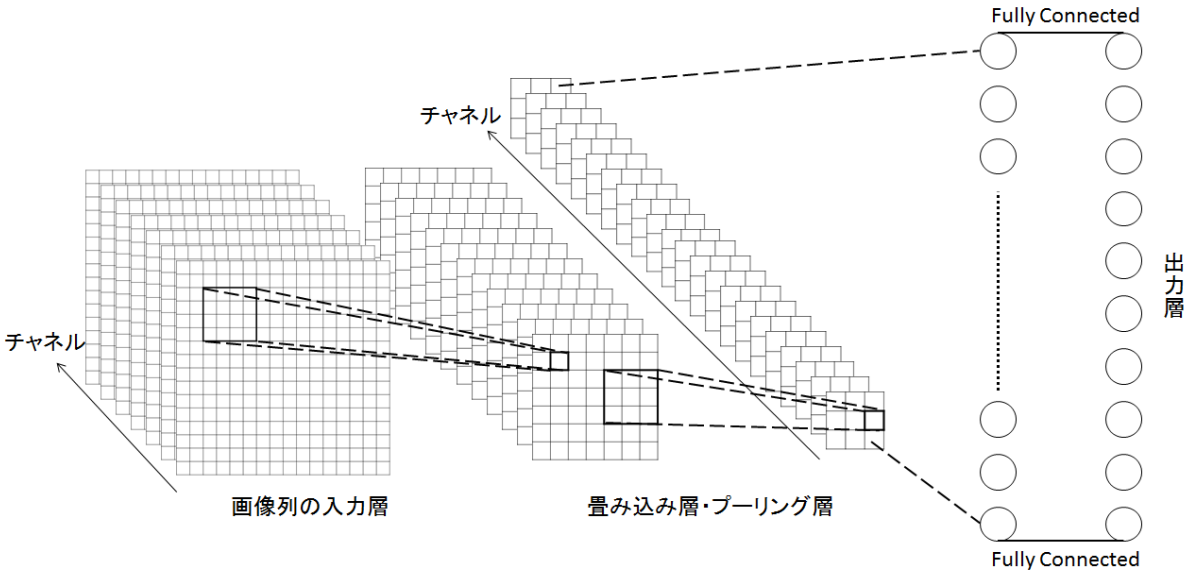

図 7: 画像列を入力とした畳み込みニューラルネットワーク

提案手法では、畳み込み層の上位層に、プーリング層を配置する。プーリング層でマックスプーリングを行うことにより、位置の違いによる影響を緩和する。

提案手法では、畳み込み層とプーリング層を上位ほど空間的なサイズが小さくなるように配置する。そして、最上位では、多チャンネルの１×１空間に畳み込まれる。これにより、二次元空間全体における動きに対する特徴量を抽出することができる。

3.2.3　出力層

出力層では、識別対象のシーン数と同数のニューロンを配置する。そして、出力層のニューロンと中間層で出力された多チャンネルの１×１の空間に対応するニューロンと全接続する。これにより、中間層で出力された二次元空間全体における特徴量と、識別先のシーンの間の対応付けを表現することが可能となる。

3.3　提案手法の動作

3.3.1　ニューラルネットワークの学習

識別させたい行動が映っているシーンを切り出した動画ファイルを用意し、そのファイルからフレーム列データを図 8 のように動画内の各時刻に対して取り出し、一つ一つのフレーム列データを学習用サンプルとして、ニューラルネットワークに入力する。

提案手法では、ニューラルネットワークの学習は、プレトレーニングとして図 9 のように段階的に階層数を増やしながらニューラルネットワークの学習を行う。 まず、入力層から 1 つめのプーリング層までのニューラルネットワークを構成し、プーリング層を出力を直接出力層に接続した、階層数の低いニューラルネットワークを構成する。そして、学習サンプルを用い、2.2.1 項で述べた誤差逆伝搬法を用いて、各層の重みを調整する。その後、畳み込み層とプーリング層を追加し、先に学習した重みを初期値として、再度学習サンプルを用いた学習を行う。この手順を繰り返すことにより、プレトレーニングと同様の効果を得ることができ、全階層の重みを適切に学習することができる。

3.3.2　ニューラルネットワークを用いた識別

学習データ投入時と同様に、動画内の各時刻に対してフレーム列を取り出し、ニューラルネットワークに入力し、出力結果を得る。出力層の各ニューロンの値は動画から切り出したシーン中の行動がそれぞれのラベルに該当する確率に相当する。そのため、当該時刻のシーンは、出力層のうち、もっとも値が大きなニューロンに対応するシーンであると判別する。

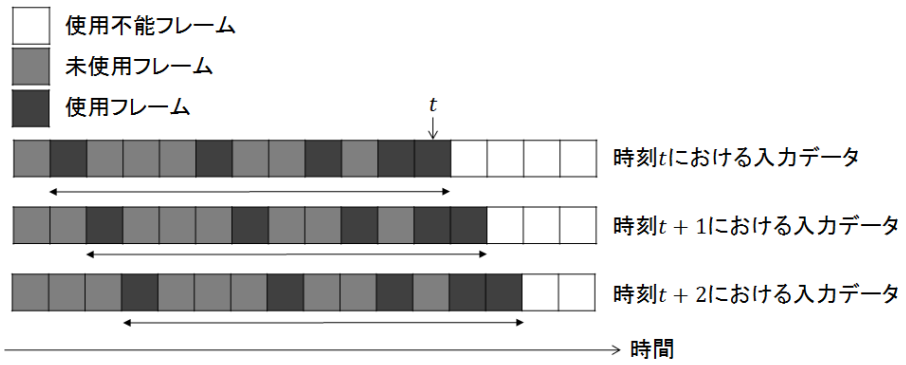

図 8: 動画からの学習用データの抽出

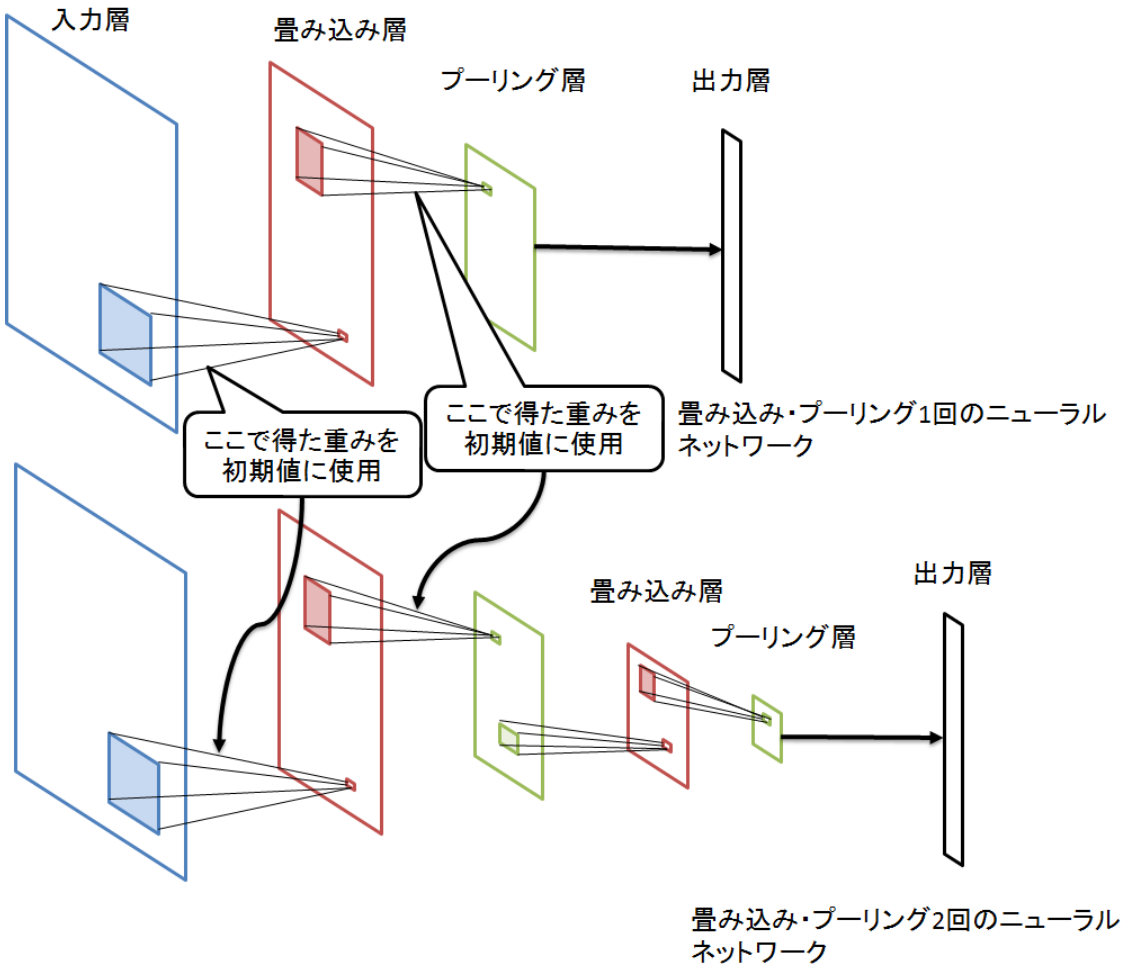

図 9: 提案手法におけるプレトレーニング

4　識別性能の評価

4.1　評価環境

4.1.1　評価に用いるデータ

人の動作について撮影した動画のデータセット [9] を用いる。このデータセットは秒間 25 フレームで撮影された動画であり、各動画の長さは 2〜5 秒である。データセットには以下の人の動作の様子が 1 動作あたり 8〜9 人分撮影されている。

- 横を向いて腰を曲げて物を拾う
- 正面を向いてジャンプしながら手足を広げたり閉じたりする
- 両足でジャンプしながら横向きに移動する
- 正面を向いてその場でジャンプする
- 横向きに走る
- 正面を向いてサイドステップする
- 片足でジャンプしながら横向きに移動する
- 横向きに歩く
- 正面を向いて片手を振る
- 正面を向いて両手を振る

本報告では、長時間撮影された動画データから、指定されたシーンを判別し、抽出することを目的としてる。そのため、単一の動画ファイル内でシーンの切り替わりが発生した際にも、切り替わった後のシーンを正しく識別できることを評価する必要がある。この評価に用いるため、本報告では、文献 [9] のデータセットをそのまま用いて識別性能の比較を行うだけでなく、文献 [9] のデータセット内の動画を結合した動画も準備する。2 つの動画を結合することにより、動画像の途中で、識別すべきシーンが変化する環境を作ることができる。そして、データセット内の動画を結合したデータセットを用い、シーンの移り変わりがある場合に、ただしく現在のシーンを認識できるのかについて確認する。

本評価では、以下の 5 種類のデータセットを学習用に用いる。

- 学習用データセット 1：手を加えていない純粋なデータセットの動画 72 個

- 学習用データセット 2：学習用データセット 1 中の同じ人物が映っている 2 つの動画の組を繋ぎ合わせてシーン切り替わり 20 フレーム前以降のみを切り出した動画 600 個
- 学習用データセット 3：学習用データセット 1 中の 60 個の動画に対し、全ての 2 つの動画の組を繋ぎ合わせてシーン切り替わり 20 フレーム前以降のみを切り出した動画 3600 個
- 学習用データセット 4：学習用データセット 1 中の同じ人物が映っている 2 つの動画の組を繋ぎ合わせてシーン切り替わり 7 フレーム前以降のみを切り出した動画 600 個
- 学習用データセット 5：学習用データセット 1 中の 60 個の動画に対し、全ての 2 つの動画の組を繋ぎ合わせてシーン切り替わり 7 フレーム前以降のみを切り出した動画 3600 個

また、テスト用には、学習用データセットに含まれない人物の動きを撮影した動画から、以下の 3 種類のデータセットを生成した。

- テスト用データセット 1：手を加えていない純粋なデータセットの学習用データセット 1 には含まれていない動画 20 個
- テスト用データセット 2：テスト用データセット 1 中の同じ人物が映っている 2 つの動画の組を繋ぎ合わせてシーン切り替わり 28 フレーム前以降のみを切り出した動画 200 個
- テスト用データセット 3：テスト用データセット 1 中の同じ人物が映っている 2 つの動画の組を繋ぎ合わせてシーン切り替わり 7 フレーム前以降のみを切り出した動画 200 個

4.1.2 提案手法に対する比較対象となる手法

提案手法　過去のフレームのうち、直近のフレームは高いサンプリングレートで、過去のフレームは低いサンプリングレートでサンプリングしたものをニューラルネットワークの入力として用いる。本評価では、現在のフレームから過去のフレームに向かって順にサンプリング間隔を 1 フレームずつ増やしながらサンプリングを 8 回行うことで、現在から 28 フレーム前までのフレームのうち、現在、1 フレーム前、3 フレーム前、6 フレーム前、10 フレーム前、15 フレーム前、21 フレーム前、28 フレーム前のフレーム、計 8 個のフレームを入力として用いる。

連続した直近 N フレームを用いる方法　ニューラルネットワークの入力として、直近 N フレームを入力として与える手法も考えられる。本評価では、現在から 28 フレーム前までの全フレームを入力として与えた場合、提案手法と入力フレーム数を揃え、直近 8 フレームを入力として用いた場合について評価を行う。

4.1.3　評価に用いたニューラルネットワーク構成

ニューラルネットワークの構造は、各層について、フィルタのサイズ、フィルタの間隔、フィルタ後のチャネル数を決めることにより、定義できる。本評価では、8 フレームを入力として用いる手法に対しては、図 10 から 13 までの 4 種類のニューラルネットワークを用いる。図 10 のニューラルネットワークは、各フレームの画像を 64 × 64 のサイズに縮小した上で入力を行う場合であり、ニューラルネットワークの入力数を削減したものである。それに対して、図 11 から 15 では、各フレームに対して 128 × 128 のサイズの画像を入力として用いる。図 11 では中間層のチャネル数が左から順に 36,36,96,96,192,512 となっている構成なのに対し、図 12 では各中間層が図 11 の倍のチャネル数を持つ構成であり、図 13 では図 12 よりさらに各中間層のチャネル数を増加させた構成である。

29 フレームを入力として用いる場合については、図 14 と図 15 の 2 種類のニューラルネットワークの構成を用いる。入力データのチャネル数が異なる以外は、図 14 は図 12 と同じ構成であり、図 15 は各中間層が図 14 の 3 倍のチャネル数を持つ構成となっている。

表 1 にこららの評価に用いるニューラルネットワークの各層におけるニューロン数をまとめる。以降、図 10〜15 のニューラルネットワークの構成に対し、入力サイズと中間層の規模から、順に 8 枚低解像度入力、8 枚入力・小層、8 枚入力・中層、 8 枚入力・大層、29 枚入力・中層、 29 枚入力・大層と呼ぶ。

表 1: 各層のニューロン数

	入力サイズ	conv1	pool1	conv2	pool2	conv3	conv4
8 枚低解像度入力	98304	69192	16200	9408	1728	384	-
8 枚入力・小層	393216	142884	34596	21600	4704	1728	512
8 枚入力・中層	393216	285768	69192	43200	9408	3456	1024
8 枚入力・大層	393216	380214	92256	57600	12544	4608	1024
29 枚入力・中層	1425408	285768	69192	43200	9408	3456	1024
29 枚入力・大層	1425408	857304	207576	129600	28224	10368	3072

4.2 評価指標の定義

提案手法、比較手法のいずれにおいても、画像データをニューラルネットワークの入力として投入した際に得られる出力層のニューロンのうち、値が最も大きなものに該当するシーンを当該時刻のシーンであると識別する。本評価では、識別されたシーンが、実際のシーンと合致しているかを評価する。評価の際には、テスト用データセット 1 を用いる場合は全テストデータ群から、100 個のデータをランダムに抽出して識別を行うのを 10 回繰り返し、テスト用データセット 2,3 を用いる場合はシーン切り替わり時を基準とした、同じ時刻に相当するデータ 200 個全てを抽出して識別を行い、以下の式で定義される精度 p で、識別の正確さを評価する。

$$p = \frac{\text{データに対して正しく識別した回数}}{\text{入力した総サンプル数}} \tag{4}$$

4.3 評価結果

4.3.1 シーンの切り替わりがない場合

まず、シーンの切り替わりが存在しない場合の識別性能について評価する。本評価では、提案手法及び連続 29 フレーム入力手法については学習用データセット 1 から抽出した学習データ 2516 個を、連続 8 フレーム入力手法については学習用データセット 1 から抽出した学習データ 4000 個を用いて学習を行い、テスト用データセット 1 から抽出したデータをテストデータとして用いて識別性能を確認した。

表 2: 行動シーンのみで学習させたときの精度

	提案手法			比較手法			
畳み込み回数	8 枚低解像度	8 枚 · 小層	8 枚 · 中層	8 枚 · 小層	8 枚 · 中層	29 枚 · 中層	29 枚 · 大層
1	57.7	52.5	40.5	42.3	45.2	50.1	36.5
2	55.4	59.7	49.5	60.9	52.2	57.4	49.2
3	59.8	62.2	53.6	60.0	51.5	58.6	48.0
4	-	63.7	57.6	60.2	49.1	62.6	51.0

表 2 より、連続 8 フレーム入力手法を除き、用いたニューラルネットワークの各層の構成によらず畳み込みの回数を増やすにつれて全体的に精度が少しずつ良くなる傾向があることがわかる。これは、畳み込み回数が増えるにつれ、より広い空間的な範囲を集約した特徴量

を抽出することができているためだと考えられる。畳み込みの回数が少ない状態ではフィルターを通しても入力解像度と比べて非常に細かい範囲に対する特徴量の列しか得られない。そのため、動画像内の全領域を通しての特徴量は捉えることができない。それに対して、畳み込み回数が増えると、より広い領域にまたがる特徴量を抽出できる。そして、最終的に畳み込みニューラルネットワークの出力が 1×1 となるまで畳み込むことにより動画像内の領域の全域にわたる特徴量を得ることができる。対して、連続 8 フレーム入力手法が畳み込み回数の増加による精度の上昇が畳み込み 3 回以降の場合で見られなかったのは、入力全体を通して連続 8 フレーム間での被写体の移動距離のような動作の範囲が小さく、2 回の畳み込みとプーリングでその範囲を十分捉えることができ、それ以上の畳み込みによる効果が得られなかったためであると考えられる。

また、小層よりも中層、中層よりも大層のニューラルネットワークの方が精度が悪化している。これは、複雑なニューラルネットワークを学習するのに十分なデータを入力として与えることができていないことが原因である。中間層のニューロン数を増やすと、より複雑なモデルを表現できるようになる。しかしながら、ニューロン数を増やした場合、ニューラルネットワークが複雑になり、学習サンプル数が十分に存在しないと、適切な学習を行うことが困難となる。

提案手法で小層のニューラルネットワークを用いた場合と、連続 29 フレームを用いる手法で中層のニューラルネットワークを用いた場合を比較すると、提案手法がより高い精度を達成できている。この原因も、連続 29 フレームを用いた中層のニューラルネットワークの方が、ニューロン数が多いため、評価に用いた学習データでは十分な学習を行うことができなかったことが原因であると考えられる。

以上の結果より、識別精度はニューラルネットワークの階層数を増やすことにより向上することができるということ、学習サンプル数が限られているシーンを識別するためには、中間層のニューロン数を抑えることにより、少ないサンプル数で十分な学習が行えるようにすることが有効であることが分かった。

4.3.2 シーンの切り替わりがある場合

次にシーンの切り替わりが存在する動画像を識別した場合の評価を行う。本評価では、提案手法、連続 29 フレームを用いる手法では、学習用データセット 2 を用いて学習した場合、学習用データセット 3 を用いて学習した場合の 2 つのパターンについて評価を行う。学習用データセット 2 からは、33040 個の学習用サンプルを抽出することができ、学習用データセット 3 からは 230640 個の学習用サンプルを抽出することができた。また、評価の際には、テスト用データセット 2 を用いた。それに対して、連続 8 フレームを用いる手法では、シー

ンの移り変わり前 8 フレーム以前のフレームの情報は、移り変わり後のシーン識別には利用されない。そのため、連続 8 フレームを用いた手法では、学習用データセット 4 と、学習用データセット 5 を学習に用いる。学習用データセット 4 からは 38440 個のサンプル、学習用データセット 5 からは 230703 個のサンプルが抽出された。そして、テスト用データセット 3 を用いて評価を行った。また、ニューラルネットワークの構成は、提案手法、連続 8 フレームを用いる手法では小層の構成を用い、連続 29 フレームを用いる手法では中層の構成を用いた。

評価結果を図 16,17 に示す。これらの図では、横軸は新たなシーンに切り替わった以後のフレーム数、縦軸はそのフレームの時点でのシーン認識の精度を示す。

提案手法の精度　図 16 は提案手法を用いたモデルが現時刻がシーン切り替わり後 27 フレーム以前の場合は 10%の精度、シーン切り替わり後 28 フレーム以降の場合に高い精度を出していることを示している。特に、学習サンプル 2 を用いた場合は、シーン切り替わり後 28 フレーム以降の入力に対して 95%以上の精度で正確に識別出来ていることが示されている。これは、ニューラルネットワークの入力として用いられるフレームが、すべてシーン切り替わり後のフレームとなるためである。シーン切り替わり後、27 フレーム目までは、ニューラルネットワークへの入力に、シーン切り替わり前のフレームが含まれる。そのシーン切り替わり前のフレームが識別結果に影響を与え、正確な識別が困難となる。しかしながら、シーン切り替わり前のフレームが入力フレームに含まれなくなると、シーンの識別に有効な情報のみを入力として用いることになり、高い精度の識別が可能となる。また、本動画像データは 25 フレーム毎秒で撮影されたものであるため、シーンの識別は、シーン切り替わり後、1 秒程度で行うことができるといえる。

本評価結果では、サンプル数が多い、学習用データセット 3 を用いて学習を行った場合の方が精度が悪化している。これは、サンプル数が増加した際に、メモリ不足のため、十分な数の学習用ミニバッチを準備できないことが原因である。そのため、この問題は、学習用のミニバッチの選択方法の工夫により解消されると考えられるが、そのようなミニバッチの選択手法は今後の課題である。

また、表 2 の場合と比べ、小層のニューラルネットワークの精度が向上している。これは、学習用データセット 1 に含まれるサンプル数が、学習用データセット 2 に含まれるサンプル数よりも多いためである。

連続 8 フレーム使用手法の精度　連続 8 フレームを用いた手法では、提案手法と異なり、入力フレーム内にシーン切り替わり前のフレームを含まない状況であっても、60%前後の精度しか達成できない。これは、連続した 8 フレームでは、動作を識別するのに十分な時間の

データが含まれていないことが原因であると考えられる。

連続 29 フレーム使用手法の精度　連続した 29 フレームを用いる手法では、提案手法と同じ範囲のフレームを入力として用いる。しかしながら、シーン切り替わり後のフレームしか入力として用いないシーン切り替わり後 29 フレーム以降についても、60%以下の精度しか達成できない。これは、(1) 提案手法と比べ、ニューラルネットワークのニューロン数が多く、十分な学習を行うために必要なサンプル数が多くなる、(2) 学習用データセット 2 には、シーン切り替わり前のフレームも多く含んでおり、学習の際にそれらのフレームが悪影響を与えているという 2 つの原因が考えられる。それに対して、提案手法では、過去のデータは、低いサンプリングレートでサンプリングされたデータしか用いないため、学習に用いたシーン切り替わりを含む各学習データに含まれるシーン切り替わり前のフレームは少ない。そのため、切り替わり前のフレームがニューラルネットワークの学習に与える影響は小さく、正確な識別が行うことができるように、ニューラルネットワークの学習を行うことができる。

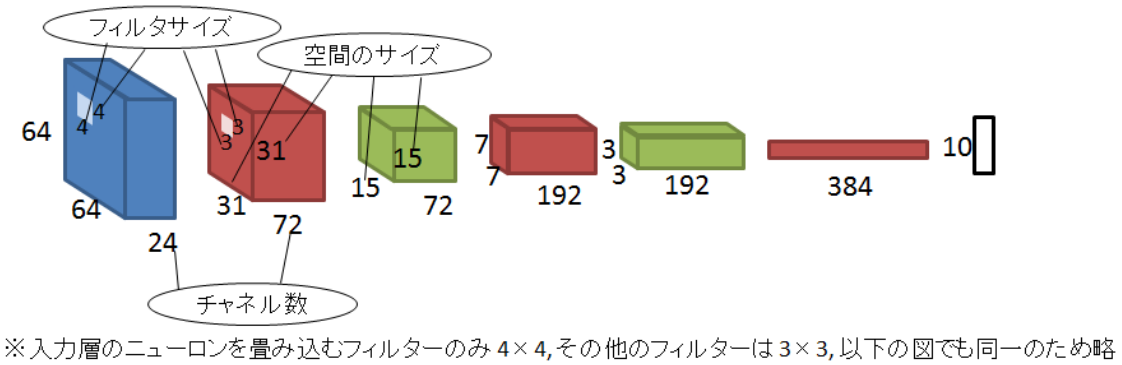

図 10: 入力サイズが 24 × 64 × 64 のニューラルネットワーク

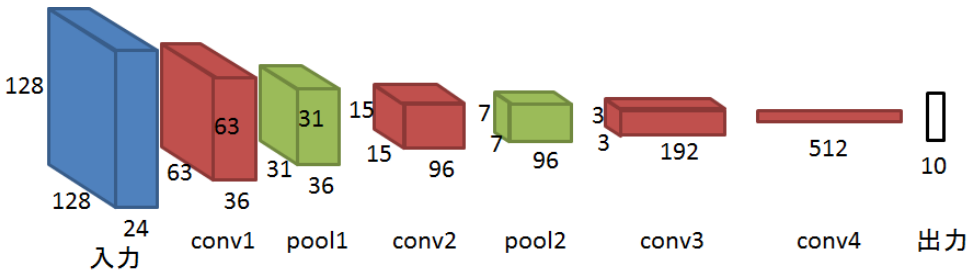

図 11: 入力サイズ 24 × 128 × 128, 中間層のチャネル数が図 12 の半分のニューラルネットワーク

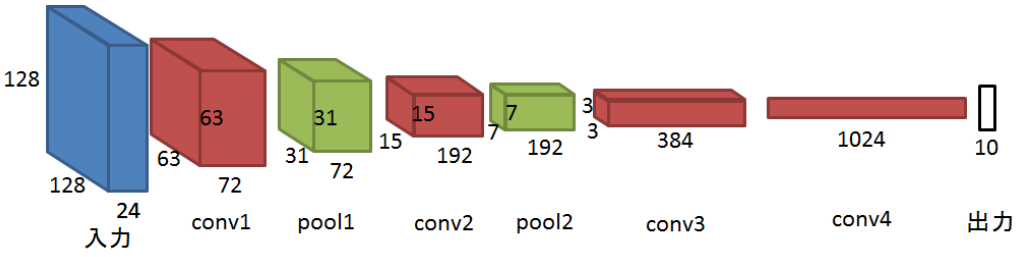

図 12: 入力サイズが 24 × 128 × 128 のニューラルネットワーク

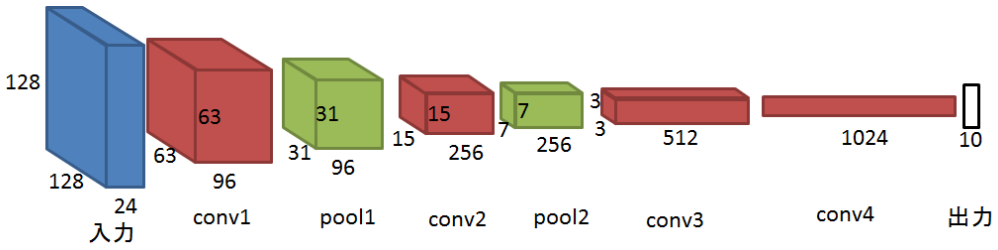

図 13: 入力サイズ 24 × 128 × 128, 中間層のチャネル数を図 12 より大きくしたニューラルネットワーク

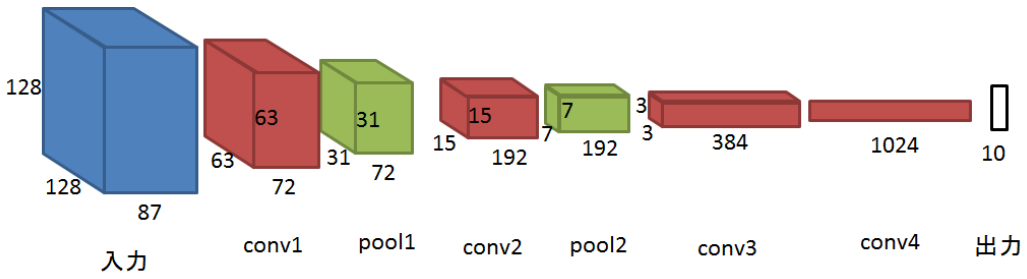

図 14: 入力サイズが 87 × 128 × 128 のニューラルネットワーク

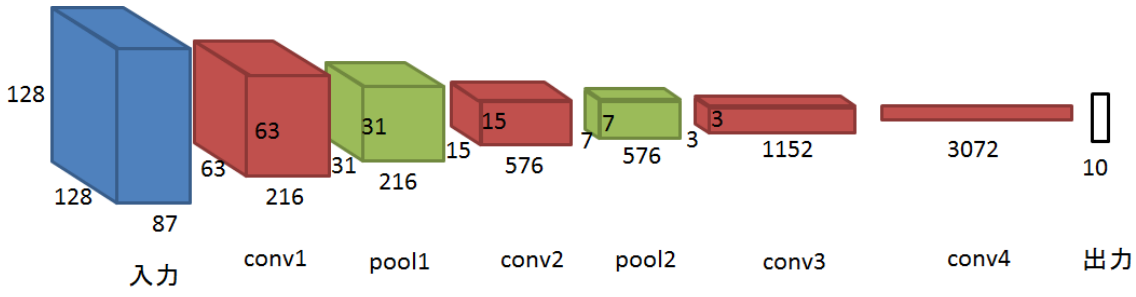

図 15: 入力サイズ 87 × 128 × 128, 中間層のチャネル数を図 14 の 3 倍にしたニューラルネットワーク

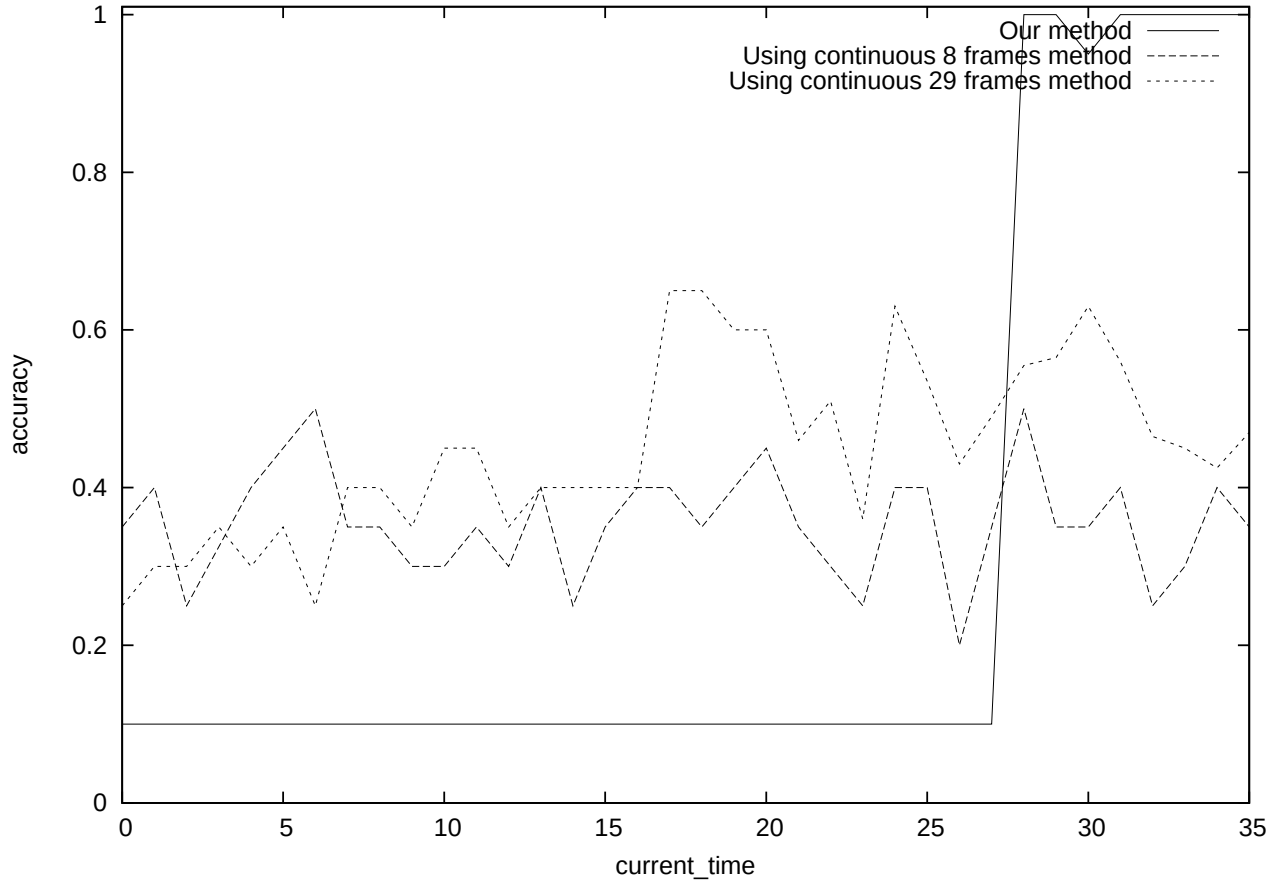

図 16: サンプル数が少ない、学習用データセット 2 または 4 を用いて学習させた場合の精度

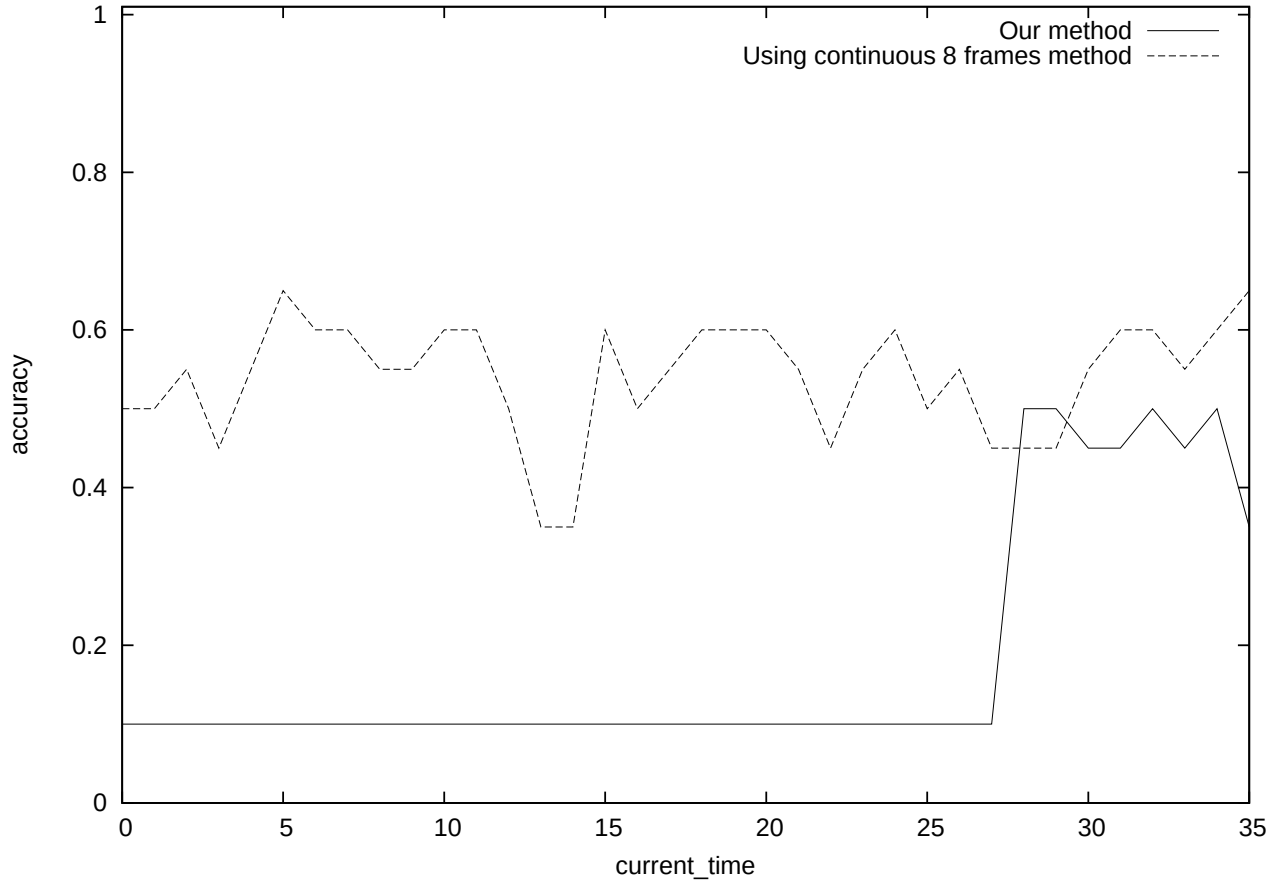

図 17: サンプル数が多い、学習用データセット 3 または 5 を用いて学習させた場合の精度

5　おわりに

本報告では、現在から N フレーム前までのフレームのうち、サンプリングした少数のフレームを入力としたニューラルネットワークを用いて、動画像のシーンを識別する手法を提案する。提案手法では、現在のシーンの識別に重要であると考えられる、直近のフレームは短い間隔でサンプリングを行い、過去にいくほどサンプリング間隔を広くする。サンプリングされた入力は、畳み込みニューラルネットワークの入力として用いる。これにより、シーンを識別するのに必要な少数の入力と出力を対応付けるニューラルネットワークを構成することができ、識別対象のシーンに対する学習データが少ない場合であっても、正確な識別が期待できる。本報告では、現在から N フレーム前までのフレームのうち、サンプリングした少数のフレームを入力としたニューラルネットワークを用いて、動画像のシーンを識別する手法を提案した。提案手法では、現在のシーンの識別に重要であると考えられる、直近のフレームは短い間隔でサンプリングを行い、過去にいくほどサンプリング間隔を広くする。サンプリングされた入力は、畳み込みニューラルネットワークの入力として用いる。これにより、シーンを識別するのに必要な少数の入力と出力を対応付けるニューラルネットワークを構成することができ、識別対象のシーンに対する学習データが少ない場合であっても、正確な識別が期待できる。

本報告では、人の動きを撮影した動画像を学習用・テスト用データとして用いた評価を行った。評価の結果、連続したフレームを入力とする学習が 50%程度の精度しか達成できないのに対し、提案手法ではシーン切り替わり後、1 秒程度で 100%の精度でのシーンの識別が可能であることが明らかとなった。

本報告では、深層学習に基づく時系列データの取り扱いとして、動画像データ中のフレーム列に対してサンプリング間隔を考慮した入力方法を用いて行動シーンの識別を行ったが、動画像データに限らずほかの時系列データについても本報告で行ったような時系列中の入力の取り方を考慮した方法を応用することができる。たとえば、ネットワーク制御を行う際には、観測されたトラヒック情報の時系列データを分析して制御することが必要であり、このような制御を分析の段階から機械的に行うために、深層学習によってトラヒック分析・ネットワーク制御の方法を学習して自動化することも考えられる。今後は、本報告のような時系列データの取り扱いの手法を応用し、ネットワーク制御のような画像の識別以外のアプリケーションへの深層学習の応用を検討する予定である。

謝辞

本報告を終えるにあたり、研究全般に関して広く御指導、御教授を頂きました大阪大学大学院情報科学研究科の村田正幸教授、ならびに研究の方針、本報告の作成に関して平素より様々な面で適切な御指導を頂きました大阪大学大学院情報科学研究科の大下裕一助教に厚く御礼申し上げます。また、研究に関して適切な御助言を多く頂きました、大阪大学大学院情報科学研究科の荒川伸一准教授、大阪大学大学院経済学研究科の小南大智助教に感謝いたします。最後に、日頃より様々な御助言と御助力を頂きました大歳達也氏、辻喜宏氏、須恵匠氏をはじめとする村田研究室の皆様に御礼申し上げます。

参考文献

[1] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[2] Z.-Q. Hong, "Algebraic feature extraction of image for recognition," *Pattern recognition*, vol. 24, no. 3, pp. 211–219, 1991.

[3] C.-L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: investigation of normalization and feature extraction techniques," *Pattern Recognition*, vol. 37, no. 2, pp. 265–279, 2004.

[4] 藤吉弘亘, "Gradient ベースの特徴抽出: Sift と hog (チュートリアル)," **情報処理学会研究報告**. *CVIM,[***コンピュータビジョンとイメージメディア***]*, vol. 2007, no. 87, pp. 211–224, 2007.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[6] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 215–223.

[7] 岡谷貴之, "ディープラーニング (技術解説)," **映像情報メディア学会誌***:* **映像情報メディア**, vol. 68, no. 6, pp. 466–471, 2014.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] "Actions as space-time shapes," http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in

Proceedings of the ACM International Conference on Multimedia. ACM, 2014, pp. 675–678, http://caffe.berkeleyvision.org/.